知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

专利文献与产业类目的映射研究

——以 2015 年度中科院专利与《战略性新兴产业分类》为例

◎ 田创 赵亚娟

中国科学院文献情报中心 北京 100190

摘要：[目的 / 意义] 提出一种基于余弦相似度的专利文献与产业类目映射模型，模型拥有准确、高效和易拓展的优点，可为后续研究提供借鉴和参考。［方法 / 过程］整理现有专利与产业类目映射方法，以 2015 年度中国科学院院所发明专利与《战略性新兴产业分类》为例，设计专利文献与产业类目映射模型并做映射实验，并根据映射成果评价模型。［结果 / 结论］专利文献与产业类目映射模型通过自然语言处理技术自动化得到专利文献与产业类目的映射组合，可实现专利到产业及产业到专利的映射，且可节省大量人力成本并方便地进行映射类目细粒度的调整，适用于大部分专利与产业类目的映射。最后，指出该模型有待完善之处，并对下一步可拓展的应用领域进行探讨。

关键词：专利文献 产业分类 类目映射 映射方法

分类号：G254.11

引用格式：田创 , 赵亚娟 . 专利文献与产业类目的映射研究 —— 以 2015 年度中科院专利与《战略性新兴产业分类》为例 [J/OL]. 知识管理论坛 , 2017, 2(1): 22-31[引用日期]. http://www.kmf.ac.cn/p/1/62/.

专利作为技术创新的成果，与产业创新水平息息相关，同时，作为参与市场竞争的重要工具，也与产业经济活动紧密联系。将专利与产业分类体系进行有效映射是技术转移和专利情报研究工作中不可或缺的步骤。

笔者在整理介绍现有专利与产业类目映射方法的基础上，制定准确、高效和易拓展的模型指导原则，以 2015 年度中国科学院（以下简称“中科院”）院所发明专利与《战略性新兴产业分类》为例，提出了一种基于余弦相似度计算的专利文献与产业类目映射模型并进行了映射实验，并根据映射成果综合评价本模型。模型通过自然语言处理技术自动化得到专利文献与产业类目的映射组合，实现专利到产业及产业到专利的映射，模型可节省大量人力成本并方便地进行映射类目细粒度的调整，适用于大部分专利与产业类目的映射。最后，指出本模型有待完善之处以及完善后可拓展的应用领域，为后续研究提供借鉴和参考。

1 现有专利与产业类目映射方法

目前，专利与产业的映射方法主要有 3

作者简介：田创（ORCID: 0000-0002-0744-9295），硕士研究生，E-mail: tianchuang@mail.las.ac.cn; 赵亚娟（ORCID: 0000-0003-3501-8131），研究员，博士，硕士生导师。

收稿日期：2016-10-09 发表日期：2017-02-08 本文责任编辑：徐健

知识管理论坛，2017（1）:22-31
DOI：10.13266/j.issn.2095-5472.2017.004

种：基于专家判定的映射方法、基于交叉检索的方法和基于概率计算的方法[1]。

基于专家判定的映射方法主要根据专家的主观判断来确立类目间的对应关系，虽然准确率较高，但费时费力，过多依赖于人工判定，不适用于大规模数据。

基于交叉检索的方法主要是用一种分类法在使用另一种分类法进行知识组织的语料库中检索，通过对检索结果所标识的类目进行分析和统计，建立两种分类法之间的映射[2-3]。该方法的局限性在于：一方面对数据量有一定的要求，如果数据量太小会造成覆盖率过低的现象；另一方面通过交叉检索得到的是一对多的映射，需要依赖统计或人工的方法进一步确定映射关系。

基于概率计算的方法将分类法类目的整体概念分解成若干足够小的单位概念，整体概念的相似度可以建立在各单位概念相似度的基础上，通过计算各小类之间的相似度从而得到整体类目之间的概率，其中小类概率之和应等于整体概率。单位概念通常由关键词表示，这样类目整体概念的相似度就转化为能够表达单位概念词的相似度之和[4-5]。该方法依据一定的规则用计算机代替人工进行语义匹配，省时省力，但映射结果还需进行一定的人工调整。

2 映射模型的指导原则

国内已有专利与产业的映射过多地依赖人工判定[6-7]，不具有普适性，且映射方法与结果均有待完善。理想的映射方法应当既满足映射的准确性，又能保证效率和可拓展性。基于此方向，设定以下指导原则：

2.1 准确性

使用专利文献中的标题与摘要信息作为专利文献的特征，使用产业类目的官方注释作为产业类目的特征，在初步分词后，提取更能精准体现专利与产业特征的动词与名词，去除不具有明显特征的停用词，以保证映射的准确性。

2.2 高效率

2012 版《战略性新兴产业分类》第三层级共有 100 个类目[8]，人工逐一对专利文献进行产业类目的映射需要大量时间，应尽可能地依托计算机技术自动化实现映射过程，减少人工的参与。本模型通过计算机编程实现快速从专利文献及产业类目中提取特征，并计算两者之间的相似度，可随着新专利、新产业分类的出现持续更新，并可以根据映射结果，高效灵活调整抽取的类目特征词词性及数量，依赖于客观的算法而不是人为的主观判断。

2.3 易拓展

本映射模型除了探讨专利文献与产业类目的映射，还将侧重于构建完善缜密的映射流程，便于下一步将映射方法拓展至其他产业分类体系及类目细粒度的调整。

2.3.1 扩展至其他产业分类

国内正在使用的产业分类除了《战略性新兴产业分类》外，还有《国民经济行业分类》《高技术产业分类》《十大重点产业分类》《统计用产品分类目录》和《产业结构调整目录》等，而现有的映射成果均基于专家判定且仅对部分产业分类进行了映射[9-10]。当新版本的产业分类发布时，重新映射往往需要大量人力物力，为避免重复工作，本模型侧重于设计完善的映射流程，通过计算机技术自动从产业类目中的官方注释抽取特征词，在保证准确性的前提下，提升可扩展性，便于映射至其他产业分类。

2.3.2 便于类目细粒度的调整

《战略性新兴产业分类》共有 3 层类目层级，第一层级 7 个类目，第二层级 30 个，第三层级 100 个，若有效利用各个类目的注释信息，可以方便地调整类目映射的层级。

3 基于余弦相似度的映射模型

基于映射模型制定的准确、高效和易拓展的指导原则，本节设计了专利文献与产业类目的映射流程，依据此流程初步实现 2015 年度中科院院所 8 309 条发明专利与《战略性新兴产业分类》第三层级 100 个类目的映射，并在第 4 节中根据映射成果综合评价本模型，映射流程如图 1 所示：

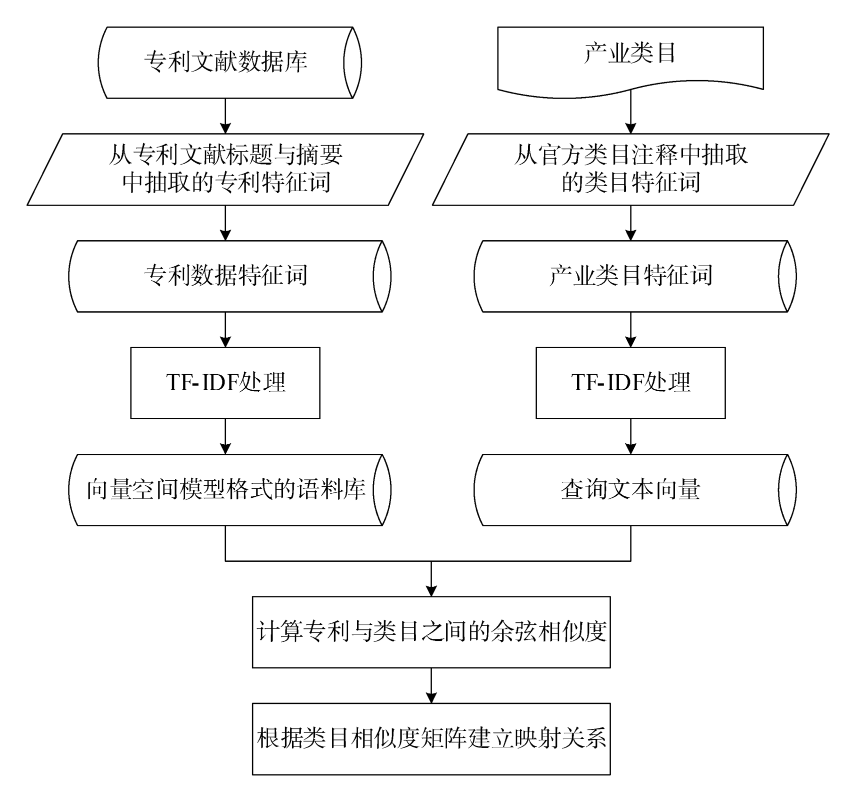


图 1 基于余弦相似度的专利文献与产业类目映射流程

3.1 获取数据

考虑到专利数据量与映射结果评价的可操作性，实验选取 2015 年度中科院各院所 8 309 条发明专利进行实验，产业类目以第三层级 100 个类目进行映射，实验共需两大类数据：

3.1.1 专利文献数据

中国科学院文献情报中心研发和维护的“中国科学院专利在线分析系统”收集了来源于国家知识产权局自 1985 年以来公布的全部中国专利信息，专利数据标准且全面，符合本文实验数据的需求。实验设定以下检索条件：①受理国家为中国；②申请日为 2015 年 1 月 1 日至 2015 年 12 月 31 日；③申请人包含中科院；④专利类型为发明与发明授权。即专利数据库中 2015 年申请人为中科院相关单位所有的中国发明专利，共检索 8 309 条，下载全部专利数据的主分类号、申请人、标题和摘要信息。

3.1.2 产业类目官方注释

《战略性新兴产业分类》包括节能环保、新一代信息技术、生物、高端装备制造、新能源、新材料、新能源汽车七大国家战略性新兴产业，用于实验的版本为“2012 版”，该版本分类表第三层级共有类目 100 个，整理其全部分类注释信息。

3.2 抽取关键词

实验基于 Python 语言，通过自然语言处理技术从原始数据中抽取关键词，共分为以下 3 个步骤：

3.2.1 分词实验

使用中文分词开源组件“结巴中文分词”[11]进行分词。该分词工具基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况，将其组成有向无环图，采用动态规划查找最大概率路径的方法，找出基于词频的最大切分组合。对于没有被收录在分词词表中的词，采用基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法，具有较好的分词能力。

3.2.2 词性判断

“结巴中文分词”工具采用的是中国科学院

知识管理论坛，2017（1）:22–31
DOI：10.13266/j.issn.2095–5472.2017.004

计算技术研究所汉语词性标记集的词性标记法，共有22大类词性，如名词、时间词、方位词和动词等。在实际操作中，发现名词和动词更能精准体现专利与产业的特征，因此在分词后进行条件判断，仅提取分词结果中的名词与动词。

3.2.3 去除停用词

停用词指在注释中类别色彩不强的中性词，例如类似、用于、提供、能够等常用高频词语，不具有明显特征，通常认为一个词在语料库中大量出现时为噪声词 [12]。为保证提取关键词的准确性，对噪声词进行删除处理，基于语料库中的词频统计与百度停用词列表确定了本实验所用停用词表 [13]。

经过以上处理后，已可批量提取较为准确的关键词，为便于观察效果，列举《战略性新兴产业分类》中第一个与最后一个类目的关键词抽取结果，如表1所示：

表1 《战略性新兴产业分类》抽取关键词结果（示例）

代码	类目名称	抽取关键词结果
1.1.1	高效节能通用设备制造	锅炉、制造、节能型、电站、锅炉、节能型、工业锅炉、节能型、船用、蒸汽锅炉、省煤器、流化床、油页岩、锅炉、秸秆、锅炉、煤泥、流化床、锅炉、蓄热、高炉、煤气、锅炉、锅炉、装置、煤粉、工业锅炉、泵、真空设备、制造、节能、节能型、真空、干燥设备、节能型、真空炉、节能型、真空、气体、压缩、节能型、制冷、压缩机、节能型、制冷、压缩机、液压、气压、动力机械、元件、制造、节能、液压、元件、制造、节能、气压、元件、制造、烘炉、熔炉、电炉、制造、节能型、炉用、燃烧器、节能型、机械、加、煤机、装置、节能、工业电炉、节能型、电热、金属、炉、节能型、辊道、窑、节能型、隧道窑、节能型、梭式、窑、节能型、推板、窑、节能型、保护、气氛、窑炉、节能型、氮化、窑、节能型、烧成、窑炉、节能型、烘烤、钢坯、步进、蓄热、加热炉、节能型、窑炉、熔炉、风机、风扇、制造、节能型、风机、节能型、工业、风扇、节能型、工业用、通风、罩、罩、制冷、空调设备、制造、节能型、工商、制冷、节能型、工商、冷藏、冷冻柜、节能型、中央空调、冷水、热泵、机组、节能型、工商、空调设备、通用设备、制造业、节能型、干燥设备
7.3.3	新能源汽车研发服务	工程、技术、新能源、汽车、电动、集成、技术、新能源、汽车、整车、技术、新能源、汽车、整车、匹配、技术、新能源、汽车、整车、轻量化、技术、新能源、汽车、整车、生产工艺、技术、新能源、汽车、整车、技术、新能源、汽车、汽车、整车、产品、质量检验、评定、新能源、汽车、电池、管理系统、集成、技术、新能源、汽车、大容量、动力电池、新能源、汽车、电池、管理系统、新能源、汽车、电机、技术、新能源、汽车、电机、制造、技术、新能源、汽车、功率、电机、开发技术、新能源、汽车电机、控制器、开发、技术、新能源、汽车、传感器、开发技术、新能源、汽车、功率、器件、技术、新能源、汽车、焊接、工艺、开发技术、电动车、传感器、电子元件、技术、动力、技术

3.3 TF-IDF 处理

为了更准确地计算类目之间的相似度，需将提取的关键词进行TF-IDF处理。TF-IDF是一种用于信息检索与数据挖掘的常用加权技术，TF（term frequency）即词频，IDF（inverse document frequency）即逆向文档频率，TF-IDF为两者的乘积 [14]。该方法用以评估某一字词对于一个语料库中的其中一份文件的重要程度，字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降 [15]。

某一特定文件中的高字词频率以及该字词在全语料库中的低频率，可以产生出较高权重的TF-IDF。因此，TF-IDF倾向于过滤常见的字词，保留具有类别区分能力的字词。

3.4 计算余弦相似度

在上一步中，已给每一个关键词都赋予了TF-IDF值，本小节将使用余弦相似度（cosine similarity）计算专利文献与产业类目之间的相似性。要计算两个文本之间的余弦相似度，步骤如下：



知识管理论坛
2017 年 第 1 期（总第 7 期）

3.4.1 向量化处理

可以将每一个专利文献的关键词与产业类目的查询词用向量来表示：

$$Patent_i=(w_{1,i}, w_{2,i}, w_{3,i},\dots, w_{t,i}) \quad (1)$$

$$Industry_q=(w_{1,q}, w_{2,q}, w_{3,q},\dots, w_{t,q}) \quad (2)$$

每一维都表示某一专利文献或产业类目中出现字词所对应的 TF-IDF 值 w，向量的维数为出现在某一专利文献或产业类目中不同字词的个数。

3.4.2 计算余弦相似度

每一个专利文献或产业类目都由对应高维度的向量表示，每个字词被赋予不同的维度，各个维度上的值为其对应的 TF-IDF 值，即每一个专利文献或产业类目已转换成一个包含字词重要程度的向量。余弦相似度通过测量两个向量夹角的余弦值来度量它们之间的相似性，余弦值越接近 1，表明夹角越接近 0 度，即两个向量越相似。因此余弦相似度可以给出专利文献与产业类目的相似度，该方法通常用于文本挖掘中的相似性比较 [16]。专利文献与产业类目的余弦相似度的计算公式如下所示：

$$\mathrm{sim}(Patent_i, Industry_q)=\frac{Patent_i \cdot Industry}{\|Patent_i\| \ \|Industry_q\|}=\frac{\sum_{j=1}^{N} w_{j,i}w_{j,q}}{\sqrt{\sum_{j=1}^{N} w_{j,i}^2}\sqrt{\sum_{j=1}^{N} w_{j,q}^2}} \quad (3)$$

3.5 类目相似度矩阵

经过上述步骤，最终可以得到 100×8 309 的专利文献与《战略性新兴产业分类》相似度矩阵，行为《战略性新兴产业分类》第三层级 100 个类目，列为 2015 年度中科院各院所全 8 309 个发明专利。

每一个专利文献与产业类目都对应相应的相似度值，正值表明正相关，正值越大表明越相关，可以根据此矩阵判定专利文献与产业类目的映射情况。同时，也可将此矩阵转置处理，用以查看每一个产业类目对应的专利文献，从而统计发现各产业类目下中科院各研究所的专利布局等信息。

4 映射结果的评价

4.1 专利文献至产业类目

4.1.1 宏观分析

在最终的相似度矩阵中，每一个专利文献与产业类目都对应相应的相似度值，实验设定相似度值大于 0 为正相关，因此对该专利文献推荐所有相似度为正的产业类目，并以相似度值大小正序排列。为便于观察整体推荐类目数量分布，作如下数据统计，8 309 条专利文献平均推荐的产业类目数信息见表 2。

专利文献与产业类目的映射频数分布直方图见图 2。

从图 2 可知频数分布直方图为右偏分布，推荐的映射数量集中在“10 至 60”，可满足一定的映射数量；进一步分析发现，映射结果可以保证对 99% 的专利文献推荐 5 个以上的产业类目，对 96% 的专利文献推荐 10 个以上的产业类目，可用于专家进一步判断选择。推荐满足率情况见表 3。

4.1.2 微观分析

在 4.1.1 小节中对映射的整体情况作了评价，本小节将深入分析具体专利文献的推荐映射成果。为保证客观合理，将选择符合映射推荐数量中下四分位数、中位数和上四分位数的第一个专利，例如专利文献与《战略性新兴产业分类》的推荐映射结果中下四分位数为 24，出现同时满足 24 个映射结果的专利有 143 个，选择 8 309 条专利中第一条满足此数目的专利，共计分析 3 个专利，对每个专利推荐相似度值排名前 5 的产业类目，详细结果如表 4 所示：

表 2 8 309 条专利文献推荐的产业类目数统计

产业分类	总类目数	平均推荐数	中位数	下四分位数	上四分位数	最小值	最大值
《战略性新兴产业分类》	100	34.5	34	24	45	0	87

知识管理论坛，2017（1）:22-31
DOI：10.13266/j.issn.2095-5472.2017.004

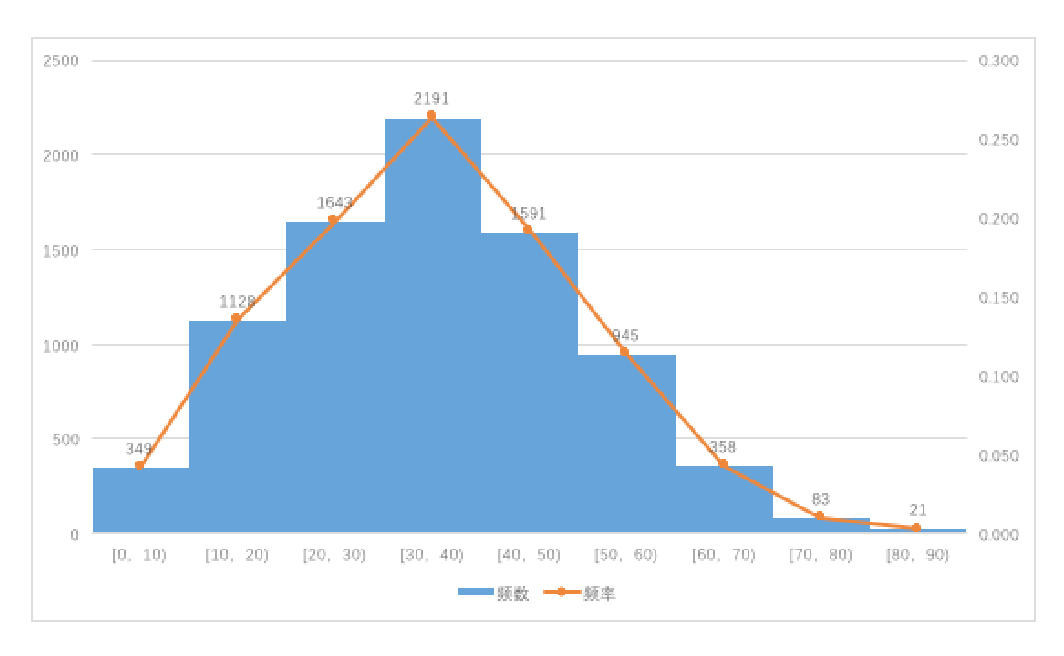


图 2 8 309 条专利文献与《战略性新兴产业分类》的映射频数分布直方图

表 3 对 8 309 条专利文献推荐产业类目满足率一览

产业分类	推荐 1 个满足率	推荐 3 个满足率	推荐 5 个满足率	推荐 10 个满足率
《战略性新兴产业分类》	99.9%	99.6%	99.0%	95.8%

表 4 专利文献与《战略性新兴产业分类》映射结果微观分析一览

选取依据	专利标题	相似度前 5 的类目	相似度值
中位数	正渗透分离方法、聚苯乙烯磺酸钠的应用及回收方法	6.1.2 新型膜材料制造	0.082 7
		1.3.2 工业固体废物、废气、废液回收和资源化利用	0.024 1
		1.4.1 节能环保科学研究	0.014 5
		6.3.1 高性能纤维复合材料制造	0.013 5
		6.4.3 智能材料制造	0.009 7
下四分位数	一种二氧化钒 / 氧化锌纳米复合粉体的制备方法	2.2.1 通信设备制造	0.085 6
		1.3.5 水资源循环利用与节水	0.085 1
		2.2.3 广播电视设备及数字视听产品制造	0.041 4
		2.2.2 高端计算机制造	0.039 2
		2.1.1 新一代移动通信网络服务	0.035 1
上四分位数	基于无线控制的超薄切片自动收集系统	6.3.2 其他高性能复合材料制造	0.019 5
		1.1.5 新型建筑材料制造	0.014 1
		1.3.5 水资源循环利用与节水	0.013 2
		6.2.3 新型合金材料制造	0.010 9
		6.2.2 高品质金属材料加工制造	0.005 9

从结果中可以看到，本模型可自动化地对每一条专利推荐一定数量的产业类目，产业类目中大部分符合专业判断，例如对专利“正渗透分离方法、聚苯乙烯磺酸钠的应用及回收方法”的推荐情况，该专利完整摘要信息为“本发明提供一种正渗透分离方法、聚苯乙烯磺酸钠的应用及回收方法。聚苯乙烯磺酸钠的应用包括，将聚苯乙烯磺酸钠应用于正渗透过程中作为汲取溶质。本发明解决了现有技术中汲取溶质普遍存在的渗透压低、反渗严重、回收困难，有毒以及与膜兼容性不好等问题”，对其推荐的 5 个《战略性新兴产业分类》类目为“6.1.2 新型膜材料制造，1.3.2 工业固体废物、废气、废液回收和资源化利用，1.4.1 节能环保科学研究，6.3.1 高性能纤维复合材料制造，6.4.3 智能材料制造”，可见推荐产业类目与专利均有一定相似性，这些推荐类目可供专家进一步判断选择，节省大量的前期人力成本。

4.2 产业类目至专利文献

4.2.1 宏观分析

将最终相似度矩阵的行与列转置处理，得到行为 2015 年度中科院各院所 8 309 个发明专利，列为《战略性新兴产业分类》第三层级 100 个类目的相似度矩阵，每一个产业类目与专利文献都对应相应的相似度值，本实验设定相似度值大于 0 为正相关，因此对该产业类目推荐所有相似度为正的专利文献，并以相似度值大小正序排列。为便于观察整体推荐专利数量分布，作如下数据统计，100 个产业类目平均推荐的 2015 年度中科院专利数信息，如表 5 所示：

表 5　100 个《战略性新兴产业分类》类目推荐的 2015 年度中科院专利数统计

专利属性	总专利数	平均推荐数	中位数	下四分位数	上四分位数	最小值	最大值
2015 年度中科院专利	8 309	386.0	356.5	229.5	540.25	8	888

产业分类与专利文献的映射频数分布直方图如图 3 所示：

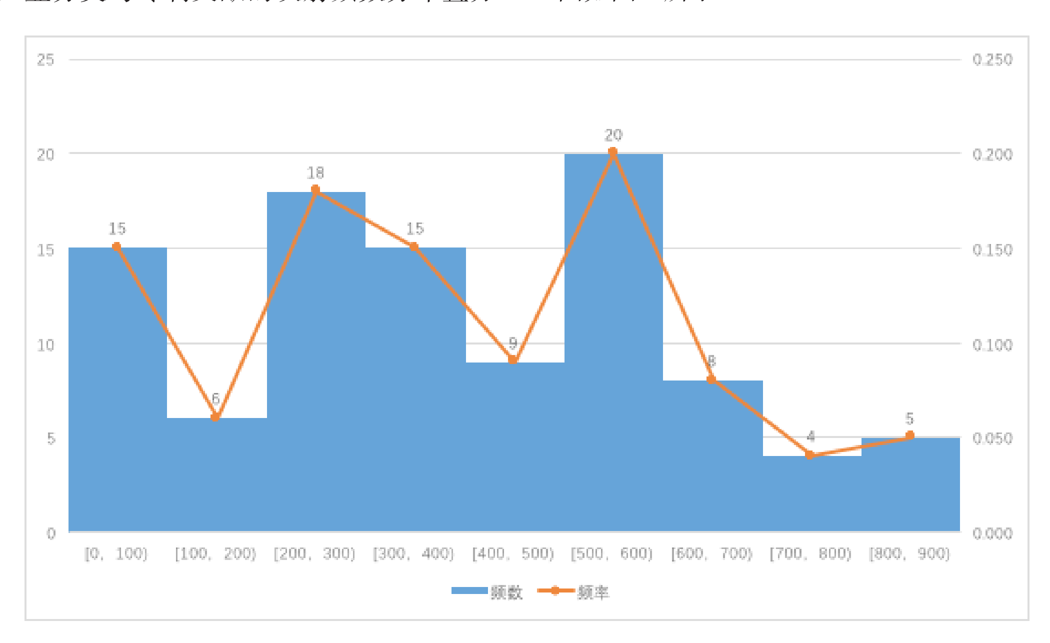


图 3　100 个《战略性新兴产业分类》与专利文献的映射频数分布直方图

可以发现，产业类目与专利文献的映射频数分布直方图没有明显的分布特征，其结果与选择映射的专利文献数据有较大关系，不同的专利数据集将有不同的频数分布，本实验映射

知识管理论坛，2017（1）:22-31
DOI：10.13266/j.issn.2095-5472.2017.004

结果中推荐的最小映射专利数量为 8，可以保证对99%的产业类目推荐10个以上的专利文献，用于统计发现各产业类目下中科院各研究所的专利布局等信息，推荐满足率情况如表 6 所示：

表 6 对 100 个《战略性新兴产业分类》类目推荐专利文献满足率一览

产业分类	推荐 1 个满足率	推荐 3 个满足率	推荐 5 个满足率	推荐 10 个满足率
《战略性新兴产业分类》	100%	100%	100%	99%

4.2.2 微观分析

4.2.1 小节中对整体情况作了评价，本小节选择《战略性新兴产业分类》的前两个分类“1.1.1 高效节能通用设备制造”和“1.1.2 高效节能专用设备制造”，从申请人角度统计具体产业类目下中科院各研究所的专利布局信息，统计数据如表 7 所示：

表 7 对 8 309 条专利文献推荐产业类目满足率一览

1.1.1 高效节能通用设备制造		1.1.2 高效节能专用设备制造	
申请人	专利数	申请人	专利数
中国科学院长春光学精密机械与物理研究所	203	中国科学院长春光学精密机械与物理研究所	188
中国科学院合肥物质科学研究院	137	中国科学院合肥物质科学研究院	150
中国科学院上海光学精密机械研究所	119	中国科学院宁波材料技术与工程研究所	138
中国科学院半导体研究所	107	中国科学院上海光学精密机械研究所	119
中国科学院宁波材料技术与工程研究所	98	中国科学院半导体研究所	114
中国科学院理化技术研究所	89	中国科学院过程工程研究所	96
中国科学院广州能源研究所	85	中国科学院理化技术研究所	93
中国科学院工程热物理研究所	73	中国科学院上海硅酸盐研究所	89
中国科学院过程工程研究所	71	中国科学院上海微系统与信息技术研究所	84
中国科学院上海技术物理研究所	71	中国科学院长春应用化学研究所	84

本模型可以从文本相似度角度观察某一产业类目下的申请人分布情况，例如2015年度，“中国科学院长春光学精密机械与物理研究所”和“中国科学院合肥物质科学研究院”在“1.1.1 高效节能通用设备制造”和“1.1.2 高效节能专用设备制造”产业领域内申请了较多的专利，在中科院各研究院所中处于领先水平。

5 结论与展望

5.1 结论

专利信息作为一种集技术、法律与经济信息于一体的战略性信息资源，是紧密联系科技和经济两大领域的纽带，若能有效利用专利数据并将其转化为专利指标去评估、监控产业发展状况，将非常有助于促进产业的快速健康发展。

笔者以 2015 年度中国科学院院所发明专利与《战略性新兴产业分类》为例，提出了一种基于余弦相似度计算的专利文献与产业类目映射模型并进行映射实验，通过自然语言处理技术自动化得到专利文献与产业类目的映射组合，实现专利到产业及产业到专利的映射，模型可节省大量人力成本并方便地进行映射类目细粒度的调整，适用于大部分专利与产业类目的映射。

5.2 改进思路

本模型虽然已可得到较可靠的结果，但准确率还有待提高，以下提供两个思路：①增加



字词的语义理解。本模型仅从字词层面进行相似度的计算，未考虑字词的语义，例如同义词、反义词等，若可以妥善处理语义层面的问题，设计算法分别对同义词与反义词设定不同的权重，将会提高模型的准确率。②专利文献增加对应专利类目的官方注释特征。本模型在计算相似度时，未考虑专利文献对应专利类目的类目特征，《国际专利分类》的官方类目注释中包含较为准确的字词信息，若能设计算法综合考虑专利文献标题摘要信息与专利所属专利类目特征，将可以进一步提升模型的准确率。

5.3 拓展应用

在妥善处理现有准确率问题后，本映射模型因有较好的扩展性，还可拓展至以下应用领域：

（1）类目细粒度的调整。本实验选用的产业类目细粒度为最小的第三层级，若方法得到进一步的完善，可以方便地调整类目映射的层级至第一层级与第二层级。

（2）其他产业分类的映射。本实验产业类目的特征词来自产业类目的官方注释信息，《国民经济行业分类》《高技术产业分类》等不同的产业分类均符合该模型的数据需求，可以调整后进行拓展。

（3）其他专利数据集的映射。本实验的专利数据为 2015 年申请人为中科院相关单位所有的中国发明专利，由于专利信息标准规范，可选择任意集合的专利进行实验，例如选择近 10 年某研究所发明专利、某大学 2015 年度发明专利和某特定领域发明专利等不同专利数据集。

笔者在下一步工作中，将继续完善本模型，尝试融合《国际专利分类》类目注释特征，以期提升映射准确性。建立完善的专利与产业映射体系，可实现专利与产业的对接，进而从产业角度出发结合科技、经济数据开展专利统计分析，对相关产业创新活动、经济发展政策决策将具有重要意义。

参考文献：

[1] 田创，赵亚娟 . 专利与产业的映射研究进展 [J]. 图书情报工作，2016, 60(1): 135-141.
[2] VERSPAGEN B, MOERGASTEL T V, SLABBERS M. MERIT concordance table: IPC-ISIC (rev. 2) [R]. Maastricht: UNU-MERIT, 1994: 1-20.
[3] SCHMOCH U, LAVILLE F, PATEL P, et al. Linking technology areas to industrial sectors: final Report to the European Commission [R]. Karlsruhe: Fraunhofer ISI, 2003: 36-52.
[4] EUROSTAT. Patent Statistics: Concordance IPC V8 – NACE REV.2[EB/OL]. [2015-09-03]. https://circabc.europa.eu/sd/a/d1475596-1568-408a-9191-426629047e31/2014-10-16-Final%20IPC_NACE2_2014.pdf.
[5] LYBBERT T J, ZOLAS N J. Getting patents and economic data to speak to each other: an ‘algorithmic links with probabilities’ approach for joint analyses of patenting and economic activity [J]. Research policy, 2014, 43(3): 530-542.
[6] 国家知识产权局 . 中国专利文献的国民经济行业分类标引工作取得阶段性成果 [EB/OL]. [2016-06-24]. http://www.sipo.gov.cn/ghfzs/zltjjb/201503/P020150325567300995160.pdf.
[7] 北京市知识产权信息中心 . 服务于产业行业的数据组织方式研究 [EB/OL]. [2016-06-24]. http://www.sipo.gov.cn/wxfw/zlxxyjjl/hgzdyj/201505/P020150827328075817725.pdf.
[8] 国家统计局 . 战略性新兴产业分类 (2012)[EB/OL]. [2016-06-24]. http://www.stats.gov.cn/zjtj/tjbz/201301/U020131021375903103360.pdf.
[9] 国家知识产权局 . 2014 年战略性新兴产业发明专利统计分析总报告 [EB/OL]. [2016-06-24]. http://www.sipo.gov.cn/tjxx/yjcg/201504/P020150422347216350682.pdf.
[10] 国家知识产权局 . 国际专利分类与国民经济行业分类参照关系表 (试用版) 编制说明 [EB/OL]. [2016-06-24]. http://www.sipo.gov.cn/tjxx/zltjjb/201512/P020151221492994057449.pdf.
[11] 结巴中文分词 [EB/OL]. [2016-06-24]. https://github.com/fxsjy/jieba.
[12] 顾益军，樊孝忠，王建华，等 . 中文停用词表的自动选取 [J]. 北京理工大学学报，2005, 25(4): 337-340.
[13] Patent_To_Industry: stopwords [EB/OL]. [2016-06-24]. https://github.com/littlewilliam/Patent_To_Industry/blob/master/StopWords.txt.
[14] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.
[15] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval [J]. Information processing & management, 1988, 24(5): 513-523.
[16] TAN P N, STEINBACH M, KUMAR V. Introduction to data mining [M]. Boston: Pearson Addison Wesley, 2006.

知识管理论坛，2017（1）:22–31
DOI：10.13266/j.issn.2095–5472.2017.004

作者贡献说明：
田　创：负责文献的搜集、整理和论文撰写；
赵亚娟：负责整体研究方向和框架把握，并对论文进行审阅和修改。

Research on Mapping Patent Document and Industrial Classification ——Mapping Between the 2015 Annual Patents of Chinese Academy of Sciences and the Classification of Strategic Emerging Industries

Tian Chuang Zhao Yajuan
National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: [Purpose/significance] This paper aims to propose a mapping model based on cosine similarity for mapping between patent documents and industrial classification. This model is accurate, efficient and scalable, which provides some references for the further research. **[Method/process]** After introducing the methods for mapping between patents and industrial classification, we designed a model for mapping between patent documents and industrial classification and completed the mapping between the 2015 annual patents of Chinese Academy of Sciences and the Classification of Strategic Emerging Industries. Then we evaluated this model according to the mapping results. **[Result/conclusion]** This model obtains the mapping results between patent documents and industrial classification automatically by the natural language processing technology and enables mapping between patents and industrial classification bi-directionally. The method saves a lot of labor costs and can easily adjust the fine-grained classification and be applied to most of the mapping between patents and industrial classification. Finally, improvements of the model are described. Some future application areas are also briefly discussed in this paper.

Keywords: patent document　industry classification　classification mapping　mapping methods